Brewster —
Against the world
opens a few times
before people catch on.
339.5 is important
BL

ComputerLetter

Volume 11, Number 23
July 17, 1995

The Index of Coincidence

Information services that collate the contents of the World Wide Web are ready to find revenue by exploring commercial models

In cryptography, the "index of coincidence" refers to the number of times a letter recurs in columnar text. For example, in two 100-character lines of English-language text, the same letter will usually appear directly above itself about seven times; the index of coincidence for English is 6.67. This kind of statistic has long been useful in cryptanalysis, the science of decoding encrypted messages, because it is a message-independent structural feature of the language that gives codebreakers a method for looking for patterns in the text.

That digression aside, the index of coincidence is a more-or-less appropriate title for a letter about the high-level indexes that Internet users consult to help them find their way through mountains of information. Looking at the origins and business plans of services such as **Infoseek**, **Lycos**, and **Yahoo**, we would have to assign them a fairly high coincidence rating. Most of them were started less than a year and a half ago, usually by university-related entrepreneurs, in many cases somewhat casually, and all are using the same raw material — the contents of the Web — to create indexes that solve the same general problem.

These coincidental beginnings are turning into equally coincidental endings as venture capitalists and larger companies snap up the information search projects one by one. Sequoia Capital has backed Stanford's Yahoo. Kleiner Perkins Caufield & Byers and Institutional Venture Partners invested in **Architext**, only a few months after it was started by recent Stanford graduates. The venture arm of mailing and marketing services firm CMG recently financed Carnegie-Mellon's Lycos. **America Online** has purchased outright a service at the University of Washington called WebCrawler, as well as wide-area search pioneer **WAIS**. Are there indexes still available? Yes. If you happen to be in the market, the WWW Worm at the University of Colorado is yet unfunded.

This turn of events interests us for several reasons. First, as investors and the originators of the search services have seen, the Web is next to useless without them. Second, this is a fine example of rendering business models in real time; those who are first to get it right have a great deal to gain. Third, because these services are becoming in effect gateways to the Web, the starting point for millions of users, they could become platforms for selling a plethora of services. In that sense, they are all candidates for the title of "online service of the future." Finally, the process of building and maintaining one of these services raises some very challenging social, technical, and intellectual-property issues that may eventually affect the way all Web-related services conduct their business.

What's a high-level Web index worth? Well, WebCrawler, with tens of thousands of regular users, brought about $1 million, up front, from AOL. The financing of Architext ($500,000 of an eventual $3 million) implies a sizable valuation. Yahoo and Lycos, it seems, should be worth far more.

(Continued on Page Two)

This Week

Net Retrieval

Continued from Page One

Search services provide a way to organize the resources stored on 50,000 servers.

Valuations are high because the indexes promise to bring some order to the geometrically growing and constantly changing Web. The Web is a mess, we're reminded by Brewster Kahle, founder of WAIS, because it's generated itself so quickly. Two years ago, there were dozens of Web servers; today there are 50,000, each spinning out hundreds or thousands of pages with no central authority, no repository, nothing to relate one site to another. The search services, in some sense, make the Web usable — and as such hold the key to much of the value of an information medium that could become the most important of the future.

And a pinch of fennel

A cynic might say it takes only a couple of mediagenic recent computer-science grads, some ad salesmen, and a spider to build such a service. A spider? Essentially all the services (Yahoo being an exception) create their structure by releasing software entities known as spiders, or robots, into the Web. A spider goes automatically from server to server, requesting Uniform Resource Locators, summary information, and in some cases whole documents, graphics, and other hypertext-linked information that might live on the server. Once this information is returned, the service can index it to create a searchable database or build some kind of hierarchical structure — that is, an index that is topically arranged and facilitates a search from general to specific information.

Greed and nostalgia

This is the strategy being followed now by several services, including **CompuServe** via its Spry acquisition, AOL via its WebCrawler acquisition, **Open Text**, one of the text-retrieval software companies mentioned in our review of text retrieval software last week ("Finders Keepers," July 10, 1995), and various universities and government labs. In fact, there are about four dozen spiders (some with other mandates) roaming the Web right now.

What all of this points out, we think, is that the commercial structure of the Internet is in a very delicate transition phase as the center of its support moves from the universities to corporations and venture capitalists.

There are three ways we know of to make money on information: by selling ad space (think controlled-circulation maga-

BUSINESS ISSUES IN TECHNOLOGY
ComputerLetter
A Technologic Publication

Editor: Richard A. Shaffer
120 Wooster Street
New York, NY 10012
Tel: 212 343-1900
Fax: 212 343-1915 • shaffer@technologicp.com

Managing Editor: John W. Wilson
P.O. Box 869
Larkspur, CA 94977-0869
Tel: 415 924-1274
Fax: 415 924-0945 • wilson@technologicp.com

Staff Editor: Brian O'Connell

Contributors: Alec Julien, Tiernan Ray, Tom Sato, Kathleen K. Wiegner

Art Director: Natalie Tilton

Production Manager: Thaddeus W. Batt

Fulfillment Coordinator: Matthew Klein

Who's Where

Private companies providing retrieval services for the World Wide Web

Company	Headquarters	Telephone	Business
Architext	Mountain View, CA	415-934-3611	Text-retrieval software and eventually a free, advertising-supported Web index
Infoseek	Santa Clara, CA	408-982-4450	Information services, one free and advertiser supported, one on a subscription and per-document basis
The Library Corporation (NlightN)	Reston, VA	703-904-1010	Information services including a free Web index with some data on a pay-per-document basis
Lycos	Pittsburgh, PA	412-268-7392	Free Web index, eventually ad supported
Open Text	Waterloo, Ontario	510-888-7111	Text-retrieval software, free Web index
Yahoo	Mountain View, CA	415-943-3231	Free information services, advertiser supported

zines), by selling information in chunks (think newspapers and newsletters), and by selling subscription-based access to information (think market-research and online services). All three models will eventually apply to the Web, and each will support different Net-related markets: a general-purpose one, maybe a middle tier, and then specialized professional services.

Getting the word out

We happen to think that sites selling advertising space will make more money, faster, than those taking other approaches: The Net has several years of rapid growth ahead of it, and the proportion of new users (the kind that are most likely to use general-purpose services) will remain high. There's also still plenty of experimental ad-budget money floating around the Net, constituting low-hanging fruit for many such sites. One potential change that should stimulate more advertiser interest is a move from the $50,000-a-quarter approach favored by the early sites to a cost-per-user approach — similar to the cost-per-thousand (CPM) model so familiar to those who advertise in print media — made possible by Web server polling, census, and analysis tools from the likes of **I/Pro** and **Digital Planet**, which make audiences more countable.

Sites selling ads are making more money than those charging users for access.

The indexes exist along a continuum of utility. There are probably two primary reasons to use the Web — entertainment and information gathering — and just because some of us have lost the ability to distinguish between these two activities doesn't mean the distinction isn't valid. We're either surfing or looking for answers. Surfers are looking for places, and places on the Net are beginning to show they can support advertising. Seekers are looking for facts or specific articles or documents, which can, as always, be sold for profit.

From the user's perspective, there's an important difference between the two search modes. If, for example, you're trying to find out the last time "electronic software distribution" was mentioned in *The Wall Street Journal*, a service that returns to you the URL for Dow Jones isn't going to be much use. On the other hand, an imaginary site that answers your question by returning the precise document you're looking for won't have much opportunity to sell advertising.

Maybe one for philately?

The architects of new business plans might also want to ask how many indexes the Net will support. The answer may depend on your favorite analogy. We imagine that the Net, at least for some time, will support fewer than ten major indexes. Why? Well, there are only nine magazines with paid circulation greater than five million. With at least three million Web users today and

Word for Word

There's more art than science in text retrieval

There are three basic ways to attack the full-text retrieval problem. First, the software can process (usually called natural-language processing) the user's query. Second, the software can apply intelligence to the way it searches through the data. Third, the software can manipulate the results that are returned before presenting them to the user. The first is to some extent a matter of religion, the second a matter of processing power (and hence cost), but the third is an area where some real competitive advantages can accrue today, as software companies figure out how best to return results to an audience more general than librarians and professional researchers.

Though this debate now concerns most directly the retrieval software vendors, the same technology will eventually have to be used by the Web-based information services to differentiate themselves from each other. For the services, text-retrieval technology comes into play once the spider has retrieved the document or URL. If the whole document has been retrieved, it is indexed — poured into a form recognizable to the retrieval engine — and in most cases discarded. If the spider has only retrieved the much-shorter URL, it is indexed as well, though this kind of database requires a less-sophisiticated retrieval engine and obviously returns less useful results.

In the future, perhaps, query-specific spiders may contain some rudimentary retrieval technology — at least the ability to look for specific words in URLs. They will not be able to parse the contents of every page for us, as this requires the creation of an index and a full-sized retrieval engine. For now, text-retrieval technical issues revolve around the user's interaction with the index created by the service.

The first challenge, that of interpreting a request, is the aspect of text retrieval that has seen the least success, in part because the problem is as complicated as language itself (what do we mean when we say what we say?). The baseline technology for retrieval requests is Boolean logic, which through a relatively unfriendly combination of ANDs, ORs, and NOTs can find with reasonable success many of the documents we're looking for.

These days, however, almost all search mechanisms provide some kind of natural-language processing. In its most rudimentary form, this entails throwing out the "stop words" in any query (in the query "When did William Faulkner get his own postage stamp?" "when," "did," "get," "his," and "own" are less-than-useful stop words) and searching for the remaining terms. This ersatz natural language is probably good enough to keep non-professional searchers happy.

maybe ten million a year from now, new services will have to work harder to get visibility and a significant number of users, though our model also suggests that there are opportunities to create more focused services. More generally the comparison suggests the medium might support three general indexes, three focused on business, and one each for fashion, sports, and technology.

Arachnophobia

A related issue is the behavior of spiders, whose automated machinations can affect the network in ways namable only by the cognoscenti but felt by all. Spiders can absorb network bandwidth, as well as overload and even crash servers with rapid-fire requests for documents. Propriety dictates that the spiders not run unattended, which pushes their activity into busier daylight hours. We can foresee some truly cosmic clashes over spider behavior. After all, it's one thing to have a few college kids amusing themselves by sending a software robot through your site, but quite another to have dozens of newly graduated entrepreneurs trying to get rich doing the same, at some expense to the rest of the community. We show how one rather advanced spider from an indexing project at the Argonne Lab does its work in the graphic on Page Six.

In more sophisticated natural-language processing, each word in a query is assigned importance based on the number of times it occurs in a document, how common it is, and the proximity of other words in the query. There are two general modes of analysis here, each of which has its partisans — statistical analysis, in which words have numerically weighted links to other words, and semantic networks based on a thesaurus or dictionary.

The first step in most modern text retrieval is building the index — in effect pouring all the documents, whether fetched by a Web spider or residing locally, into a meat grinder that renders them recognizable to queries. For this computing-intensive task one company uses supercomputers, another uses hundreds of PCs linked together over a fiber-optic network. Part of the craft is in making these indexes as small as possible. There is also a move afoot (led by, of all unlikely institutions, the U.S. Air Force) to make the indexes interoperable between vendors.

Semantic networks don't require this kind of indexing but do need tremendous processing horsepower on the back end. These systems, provided only by **ConQuest** and soon by **Oracle**, tend to work better in environments with a familiar collection of documents, and are less likely to miss relevant documents that may not contain precise search terms. Statistical analysis, however, which in effect builds a thesaurus by looking at the documents in question, is easier to maintain and in general gives better results when searching in rapidly changing environments. Both technologies are applied to organizing the results of the query in step three.

The second part of the text-retrieval problem, processing the query, has largely been solved. All the retrieval-software companies to a greater or lesser degree have managed to implement fast search algorithms (**Open Text** claims to be fastest), can distribute processing over multiple processors, and help users maintain the databases with a minimum of effort.

The third step, organizing the results of a query in a useful manner, is to a large extent a user-interface issue. The role of the software is to help the user quickly decide what kind of documents she wants to see more of, and allow her to rapidly and easily inflect her search criteria. This feedback loop, because it can be a client-side process, benefits most from the appearance of more powerful desktop machines. **Architext** has handily implemented an algorithm for creating summaries without benefit of a semantic network; **Verity** was for a time interested in acquiring or licensing that technology but will instead implement its own version. Some speculate that it's not information *in* the document but *about* the document (context, reviews, information about popularity and usage of the document) that is the key to better retrieval.

Other benefits that client software can provide include relevance ranking, in which more relevant documents appear at the top of the results list, and clustering, in which related topics appear under appropriate headings. Whatever technical choices they make, the key to the future for these companies is to keep their software flexible enough to bolt on new capabilities as they appear and as the amount of processing power available at both the server and desktop level increases. □

Although we're led to believe that it's not easy to get a spider working correctly, there is some genuine concern about what might happen if the number of spiders continues to increase. One alarming scenario is of the personal spider that could be sent out by individual users, a sort-of do-it-yourself Web indexing kit. An even more alarming sub-scenario is that of the spider-canceling spider, patrolling the Web to block spider-generated requests, possibly turning the entire Net into an arachnid battleground.

That said, we'd also argue that certain kinds of spiders could become ideal database mangers, and if run nightly (along with backup, say) could make internal Web servers an extraordinarily cost-effective way to manage documents inside a corporation. This eventuality could give the indexers or other owners of domesticated spiders a standalone software product to sell, as well as provide more ammunition for those who push the Web as a potential competitor to **Lotus Development**'s Notes and other groupware applications.

Ownership, authority

Some of the intellectual-property issues that the Web indexes raise are or should be easily solved. For over a year there has been a standard protocol for excluding

Rebuilding the Web

One version of spider-created retrieval resources

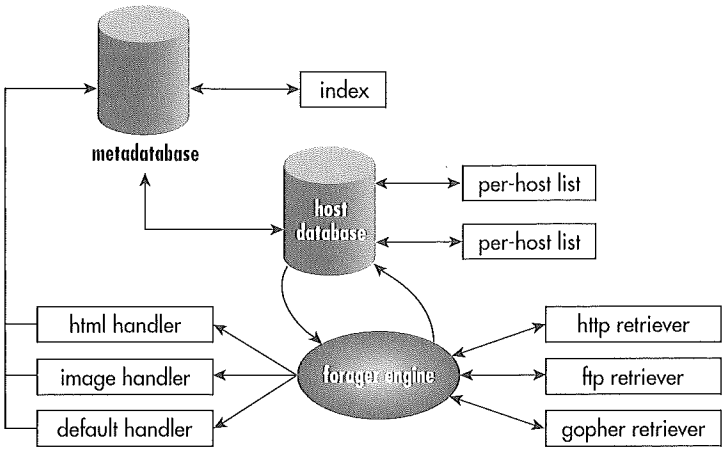

Diagram of an experimental system to create a searchable index and database of Web-based resources. The Web Forager designed by Argonne National Lab is more elaborate than most of those discussed in this letter in that 1) it retrieves not just HTML text but gopher and FTP documents, and 2) it locally caches some of the documents it retrieves. The "forager engine" is in effect the spider. The system acquires documents from the Internet with the "retrievers" to the right of the spider, indexes them with the "handlers" to the left of the spider, and caches documents in the "host database," which notes which documents came from which sites in its "per-host lists." Both the cached pages and indexed documents can then be retrieved using text-retrieval software.
Source: Argonne National Lab

The important issue facing spider owners is how to turn their collections of information into viable businesses.

robots and spiders from your server — although not all spiders yet follow it. Another possibility is a development similar to the United Way — a charitable organization that was formed essentially to keep a dozen canvassers from showing up at your door every day. A super-spider (Lycos, it seems, is in the best position) could become the authoritative index. This development would allow sites to maintain greater control over how and where their site was indexed and distributed by admitting a single spider. For the protective types, it would mean only one to turn away. Even in a world of consolidated giving, however, large and credible charities such as the American Cancer Society continue to raise funds on their own, as most likely some existing indexes would maintain their own spider.

On the other hand, this is not a closed argument. A text search (see again, last week's letter) requires far more processing power than a request for documents, so a query that is exploded to hundreds or thousands of servers would affect these machines more dramatically than would a gradual indexing of their contents. In addition, a badly formed query sent to several dissimilar servers could return very little of interest.

Getting a piece of it

There are a couple of possible alternatives to indexing the entire contents of the Web. One is a protocol known as Z39.50, championed by WAIS, which allows users to query several servers at once and combine and rank the results. The obstacle is that not all text or Web servers are compatible with this process. Two other experimental projects, known as Harvest and Aliweb (they may be related, we're not quite sure) have been designed to try to solve some of the problems that might be caused by multiple spiders through automating the generation and collection of meta-information.

Another possibility is the query-specific spider, perhaps one following the instruction to "go get me all the information about healthcare." The computer science department at Stanford University is currently testing such an approach; spiders return with preliminary results in 24 hours, the user is asked to refine her criteria, and then the spider returns again a day later with more specific results. This, again, could cause bandwidth and server problems. The tradeoff with spiders is between speed and responsibility; they can request many documents in a short period of time (and thus affect servers) or take longer to run.

These issues need to be solved. A year or two from now, the Web will be too big and still changing too quickly for spiders to be a viable way to encapsulate it. At present, services such as Lycos re-index weekly. In a Web universe four times bigger, this process will take a month — meaning the information in the index is quite likely to be uselessly out of date.

Selling the beast

The owners of the spiders, of course, are well aware of these issues; more pressing for them, perhaps, is how they can turn their comprehensive collections of information about information into a business. The obvious answer: advertising. In fact,

an index of the Web is general enough to support advertising. There may not, however, be many collections of documents that can. Some sites, for example an index of all the molecular biology-related Web sites, wouldn't tempt broadly-focused advertisers but might well be of interest to pharmaceutical companies or makers of gene-splicing equipment. Other collections of documents — **Bell & Howell** subsidiary UMI (formerly University Microfilms) for example has some documents that cost $250 each — wouldn't attract enough users to support advertising. It's the same with magazines; populist publications rake in the ad dollars while scientific journals support themselves with subscription revenue.

Yahoo may be both the most interesting and the most challenged of the Web indexes.

Just as these publications serve different audiences with different approaches, the following representative services are pursuing their own paths to creating profitable markets for their services.

- Infoseek is pursuing a hybrid strategy with both a free and for-pay service. From its five advertisers Infoseek gets 1.5 cents for every user that comes into the free area, a cost-per-thousand of $15, about the same as monthly computer publications. If Web advertising does anything, however, it will make the concept of cost-per-thousand obsolete; as Web-publishers are able to target audiences more precisely, the price per exposure should certainly rise. On the non-gratis front, Infoseek's service, at $9.95-a-month, promises to return more than ten hits on a search and has a more extensive collection of documents than the free service. Founder Steve Kirsch says he makes a lot more money from advertisers on the free services than he does from subscriber fees, which explains his current emphasis on building a market of "casual" Net users and making Infoseek a service they can't live without. Infoseek's experience with the fee-based service is also an indication that the ad-supported model is the right approach for now.
- Yahoo is perhaps both the most interesting and the most challenged of the Web indexes. Because it's free and extraordinarily easy to use, it will continue to attract new users. Yahoo's hierarchical approach also means there is an almost infinite number of special-interest sections to sponsor. We think Yahoo is challenged because it may need to find a way to maintain its site that doesn't depend on manual categorization or user submissions. Its variety — thousands of increasingly refined topic-specific areas — makes an interesting contrast, for example, to WebCrawler, which presents the user only a single-screen interface. Yahoo is a site, whereas the WebCrawler is a service. Look for a coming redesign at Yahoo that will make the site more friendly to advertisers.
- Lycos, the CMU index, will use its venture capital infusion to pursue yet another approach to profitability. For now most of Lycos' revenues will come

from licensing the use of its index to other services, for an up-front fee and a yearly charge. **Microsoft** has licensed the index for use on the Microsoft Network, as has an indexing subsidiary of **The Library Corporation**, NlightN. Several other deals are pending. Interestingly, all services pay the same price, though presumably MSN will generate more traffic than others. The plan eventually is for advertising at the home site to generate significant revenues as well.

- Architext, a startup in Mountain View, Calif., that has gotten some good publicity lately, will take a three-part approach: an ad-supported Web-based service, a standalone software product designed for Web servers, and a full-fledged text-retrieval engine. As we understand it, Architext's proprietary text retrieval is refined to reveal documents based on simple queries (the assumption being that most people try to find what they're looking for with a single word.) The company also has some clustering capabilities for returned documents, so that a query for Napoleon would return, for example, clusters of documents under the heading Napoleon Bonaparte, Napoleon III, and Ross Perot.
- NlightN is taking yet another approach to selling information. In addition to licensing the Lycos index, the service has used its own search-and-retrieval software to index hundreds of public-domain and proprietary databases that its parent company has historically administered for libraries and other organizations. Web hits are free, and documents from the other databases generally cost a dime, though some cost far more. To pay for their hits users deposit money with NlightN via credit card before they can retrieve. While there's something strange about NlightN (the bizarre name? the odd look of the Web site?), this is a business model referred to most often by others not yet in the business.

Those able to build an audience today will be in a position to sell new services in the future.

Metamarketing

There should be room in the ever-expanding Web for at least the index services we've mentioned. All told, there are close to 100 sites that offer some sort of searching or indexing. We can't think of many reasons why a site wouldn't want to be indexed in as many places as possible — at least the vast majority of sites that contain nothing more than marketing information. Eventually, some sites might pay to be listed in the right place — how much, for example, could **Netscape Communications** charge for the right to be named the "cool site of the day?" It's this kind of service (and doubtless others), based simply on having the right traffic, that some of these indexes might want to pursue. One doesn't have to be a codebreaker to realize that those who are able to build an audience today — regardless of changes in the technology, size, and social contracts of the Web — will be in the best position to sell such new services in the future. □

At Random:

Serving up a buggy soup

Online software support is a boon for everyone, right? Vendors save money, and e-mail-connected users get better support. Maybe not. We're experiencing a more-or-less common problem that the new **Microsoft** Office seems to cause on Power Macs. Toward solving it, we e-mailed our plaint to Microsoft support on three different online services, hoping to be trebly reassured. Support answer one: A simple reinstallation should solve your problem. Support answer two: We're aware of the problem and working on a patch to correct it. Support answer three: Your problem is possibly a result of a conflict with **Now Software**'s Now Utilities. We elected to believe none of these answers, and are, at present, seeking a fourth opinion. Meanwhile, the online support folks might want to consider centralizing their facilities. □